PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DHANKAWADI, PUNE - 43
**BE PROJECT SYNOPSIS**

DEPARTMENT OF COMPUTER ENGINEERING ACADEMIC YEAR : 2019-20

**Group Id** : 35

**Project Group Members**:

| Roll Number | Name | Guide |
|---|---|---|
| **4235** | Hrushabh Hirudkar | 1. Dr AS Ghotkar<br>2. Dr SS Sonawne<br>3. Dr GV Kale |
| **4242** | Anuj Kanetkar | |
| **4256** | Shriniwas Nayak | |

**Project Title** : Generating relevant questions for a query using natural language processing techniques

**Domain** : Artificial Intelligence/Machine Learning

**Sponsorship** : Tech Mahindra

**Department** : Maker's Lab

**External Guide** : Mr Nikhil Malhotra

**Synopsis** :



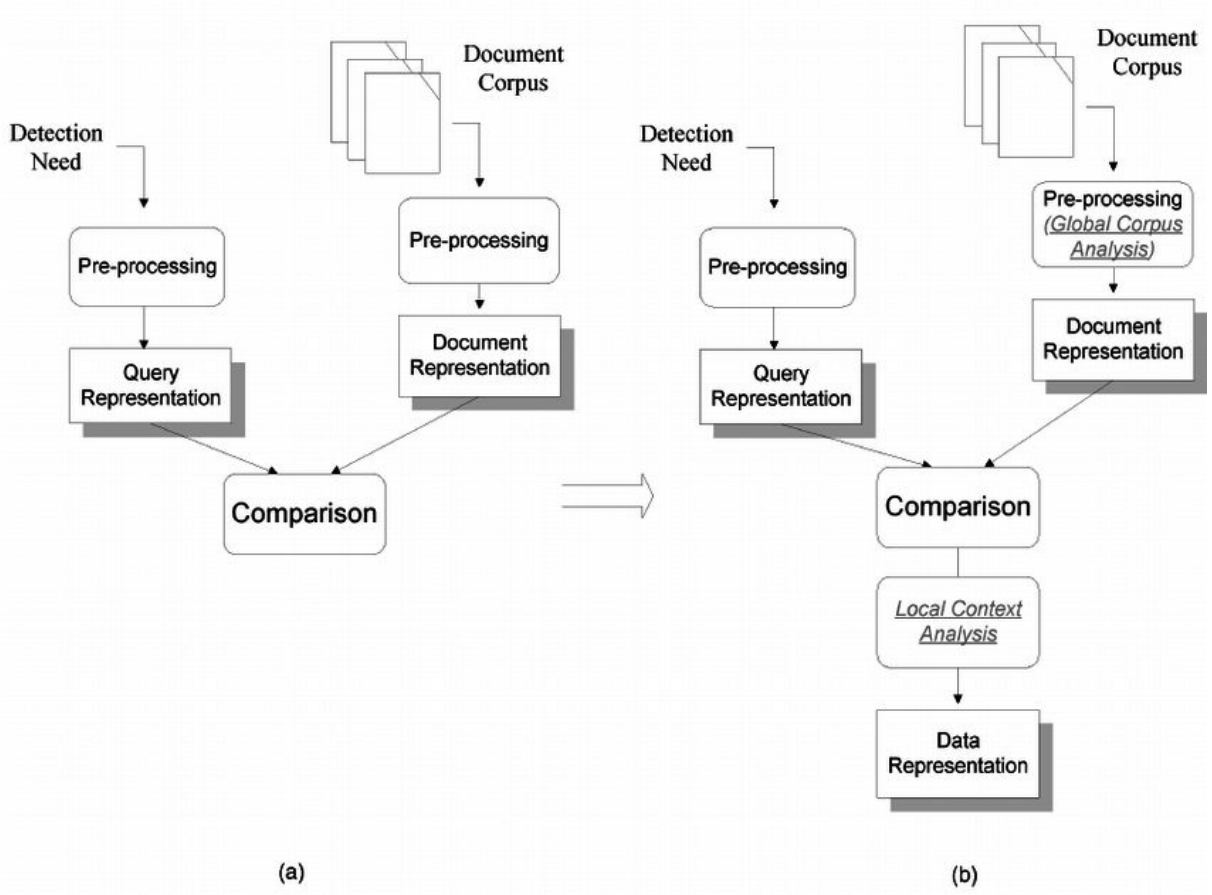
## Abstract :


In today's technologically advancing world, getting answers for questions, general or specific is germane to the development of the user and as a result, development of the overall community. In this synopsis, we propose to find the relevant questions to a query entered by the user to solve doubts. Using text similarity, we can find out relevant questions to the query put forth by the user. This synopsis discusses the approaches for text similarity for questions and the use of different metrics to evaluate the similarity between questions.


## Keywords:


NLP, Word2vec, Information Retrieval, Lexical Similarity, Semantic Similarity, String Similarity, Corpus Similarity

**Architecture :**




**Mathematical Model:**

Cosine Similarity is one of the methods used to calculate the similarity between two documents ‘A’ and ‘B’. Each document consists of different words which are represented as two different vectors and the cosine product helps to find the angle between them, thus helping us to conclude on their similarity.

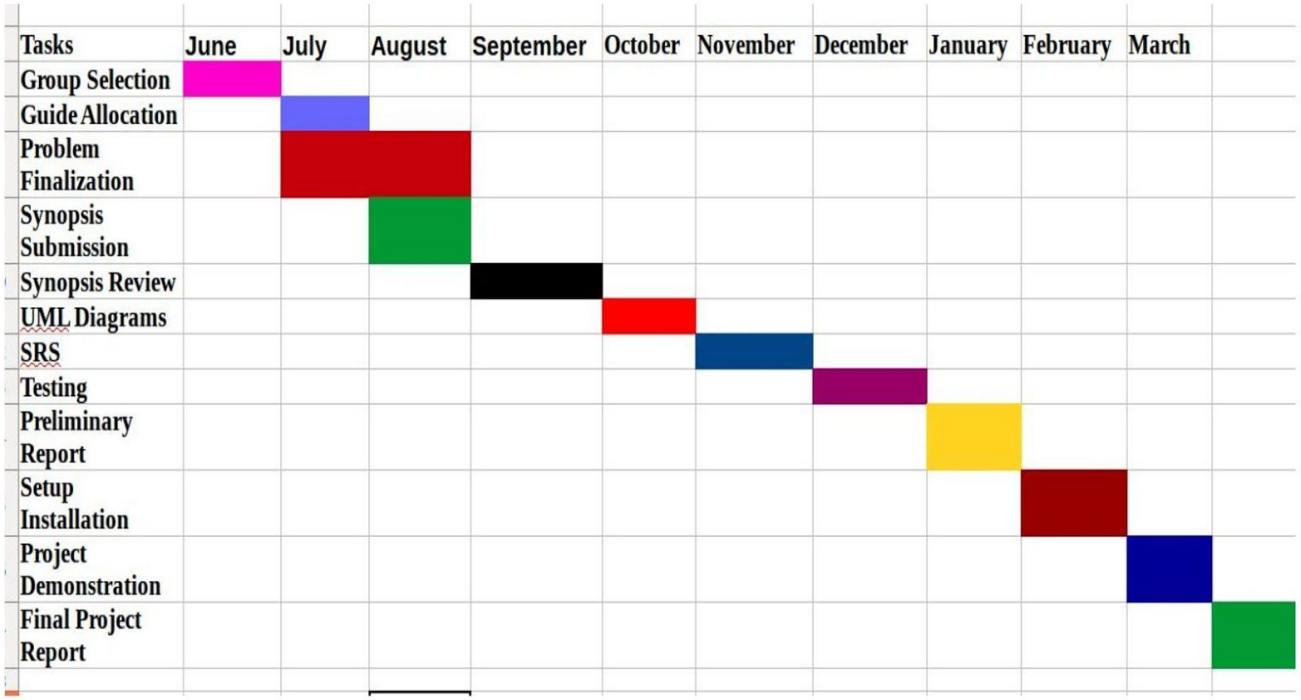
For example, If the cosine of the angle between vector of the two documents is 0 then they are similar and if it is -1 it represents that they are completely different.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

**Name of Conferences where paper can be published:**

1. IEEE International Conference on Machine Learning and Applications (ICMLA) June 2020
2. ACM International Conference on Machine Learning and Computing (ICMLC) - February 2020

**Project Time Line**




**References:**


1. Song Y., Roth D. (2015). "Unsupervised Sparse Vector Densification for Short Text Similarity", Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. DOI:10.3115/v1/N15-1138

2. Sidorov et.al.(2015). "Computing text similarity using Tree Edit Distance", 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC). DOI: 10.1109/NAFIPS-WConSC.2015.7284129

3. Kashyap, A., Han, L., Yus, R. et al. Lang Resources & Evaluation (2016) 50: 125. https://doi.org/10.1007/s10579-015-9319-2

4. Schwarz C. (2019). lsemantica: “A command for text similarity based on latent semantic analysis”.

5. Gomaa W, Fahmi A. (2013). “A Survey of Text Similarity Approaches”, International Journal of Computer Applications (0975 –8887)Volume 68–No.13